In the wake of gene-oriented data analysis in large-scale bioinformatics studies, focus in research is currently shifting towards the analysis of the functional association of genes, namely the metabolic pathways in which genes participate. The goal of this paper is to attempt to identify the core genes in a specific pathway, based on a user-defined selection of genomes. To this end, a novel algorithm has been developed that uses data from the KEGG database, and through the application of the MCL clustering algorithm, identifies clusters that correspond to different “layers” of genes, either on a phylogenetic or a functional level. The algorithm’s complexity, evaluated experimentally, is presented and the results on two characteristic case studies are discussed.

Keywords: Bioinformatics; metabolic pathways, clustering algorithm; phylogenetic analysis.

1. Introduction

1.1. Metabolomics

Metabolomics is the scientific study of the chemical processes that involve metabolites. Specifically, metabolomics is the “systematic study of the unique chemical fingerprints
that are left behind specific cellular processes” [1]. The metabolome represents the collection of all reactants (enzymes, proteins or other chemical compounds) in a biological cell, tissue, organ or organism, which are the end products of cellular processes. Thus, while mRNA gene expression data and proteomic analyses do not tell the whole story of what might be happening in a cell, metabolic profiling can give an instantaneous snapshot of the physiology of that cell. One of the challenges of systems biology and functional genomics is to integrate proteomic, transcriptomic, and metabolomic information to give a more complete picture of living organisms.

1.2. **Metabolic Pathways**

In biochemistry, a metabolic pathway is a series of chemical reactions occurring within a cell where a principal chemical is modified towards an end product. Enzymes catalyze these reactions, and often require dietary minerals, vitamins, and other cofactors in order to function properly. Because of the many chemicals (i.e. metabolites) that may be involved, metabolic pathways can be quite elaborate. In addition, numerous distinct pathways co-exist within a cell. This collection of pathways is called the metabolic network.

For the current study, pathway-related information was retrieved from KEGG Pathway Database (KEGG) [2]. KEGG is a collection of manually drawn pathway maps representing our knowledge on the molecular interaction and reaction networks for various processes and systems including “Metabolism”, “Human Diseases” and “Cellular Processes” among others.

1.3. **Motivation**

The analysis of completely sequenced genomes can yield useful insight into the evolution and multi-level organization of organisms. From the lowest level of molecular organization where individual enzymes are formed, to the higher level, where sets of enzymes group into metabolic networks, the effects of gene evolution can be detected. With the current advances in genomics and proteomics, it has become imperative to explore its impact as reflected in the metabolic signature of each genome [3]. To this end a methodology is presented, which applies a clustering algorithm to genes from different species participating in the same pathway.

1.4. **Structure**

The general goal of the current study is analyzed in section 2 and several similar approaches towards the same direction are discussed. In the next section, the main algorithm process is presented and explained in depth, evaluating at the same time its overall complexity and efficiency. Moreover, the results of two significant test cases among the various cases that were tested are presented, demonstrating the effectiveness of the algorithm. Finally, with regard to the aforementioned results, the paper closes with a discussion on the perspectives for further efforts and studies.
2. Problem Description

Considering the fact that the majority of the known genomes from all three domains (Archaea, Bacteria and Eukaryota) use highly similar chemical processes in the form of metabolic pathways, it is a logical step to assume that the dissimilarities in the particular expressions in the metabolic pathways of different genomes can be viewed as the result of the evolution of a reference pathway existent at the common ancestor of those genomes. Based on this concept, if each pathway is represented by a graph model, then starting from the graph of the reference pathway, the current pathway formations have essentially emerged after the addition or deletion of certain nodes and/or edges from the reference graph, while preserving the main body of the pathway virtually unchanged. In fact, by observing the graph representing a metabolic pathway using several genomes as a reference, several compact sets of edges (i.e. reactions) can be distinguished that seem to be highly conserved in the genomes. At the same time there exist several sub-graphs that have been added or deleted, making evident the differentiation of the pathway expression between the genomes. This observation leads us to the conclusion that the transformations that have been carried out in the gene-level of the genomes through evolution due to mutations of the genome sequences have resulted to changes in the topology of the metabolic pathway while retaining the overall process.

Thus, the objective of the current study is to extract a meaningful clustering of the genes participating in a common metabolic pathway of several genomes by comparing the genes sequences and by evaluating the degree of homology between genes from different genomes. The result of the applied process is the extraction of cohesive gene groups that can give us information about the evolutionary similarity between the genomes examined or the uniqueness of some reactions in particular genomes. This knowledge can then be used in order to correlate some macroscopic differences between the genomes with the differences that arise between them in the pathway level.

2.1. Related work

Genes usually do not act individually but form functional or structure organizations, exemplified by metabolic pathways. As metabolic pathways are essential to the survival of organisms, and their evolution has been under debate for more than half a century [4], a combined phylogenetic and phenetic analysis of pathway topology might expand the understanding of the evolutionary processes molding their form and structure.

Several groups have carried out phylogenetic analyses based on metabolic pathways, deriving phylogenetic trees from the information of individual pathways [5-7], the presence and absence of entire pathways [8], or the reaction content of entire pathways [9]. These studies have provided valuable insight into the evolution of metabolism; however, as phylogenetic trees, they have generally diverged substantially from trees based on 16S rRNA, the most used molecule for phylogeny construction. A common feature of phylogenetic trees based on metabolic information is that, owing to similar evolutionary pressures, organisms in similar habitats tend to be clustered together, and Aguilar et al. [10] therefore regarded such trees as phenetic rather than phylogenetic.
Fig. 1. Flowchart demonstrating the steps of the algorithm.
Furthermore, one group showed that trees based on different subsets of metabolic networks were different [10], and another result also indicated a similar situation when several different pathways were used to construct trees separately [7].

On the other hand, phylogenetic profiles are commonly used in evolutionary studies, as they are based on sequence similarity. There are several recent approaches that either directly utilize phylogenetic profiles for functional prediction of gene clusters [11] or combine them with other biological data sources for increased sensitivity [12]. A slightly different approach, and an intermediate step towards the work presented in this paper, is to produce a tree-like structure of gene clusters in order to reconstruct the evolutionary relationships between them [13]. However, it is shown that the output of large-scale reconstructions is notably more difficult to interpret biologically. Our approach extends this work by aiming to extract close gene associations from metabolic pathways through unsupervised clustering at a sequence level. This level of association can be enhanced if the phylogenetic relationship of the corresponding genomes is taken under consideration. Although there are several approaches in literature addressing the issue of projecting the underlying evolutionary process on the metabolic pathways [14] or identifying multiple branches to the same metabolic goal [15], the presented methodology relies on sequence similarity to provide an estimate of pathway conservation. Finally, in order to achieve the same level of user interaction as other existing algorithms [16], the methodology also provides for a visual output of the process by setting up a color scheme on the popular KEGG metabolic maps.

3. Algorithm overview

The algorithm comprises four main steps; (a) the extraction of protein sequences from the different species under study that participate in a specific metabolic pathway, (b) the construction of the homology matrix of the retrieved sequences, (c) the clustering of the matrix into groups of proteins based on a defined similarity metric, and (d) the calculation of the estimated phylogenetic distances of the species based on the produced clusters.

The algorithm accepts as input a KEGG pathway map identifier mapID, and a list of n genome identifiers that comprise the target data set. For each run a clustering algorithm is selected, e.g. the Markov Cluster Algorithm (MCL) [17] or the Expectation-Maximization Algorithm (EM) [18]. Moreover, its algorithm-specific parameters are set, such as the inflation parameter for MCL. Finally, a parameter (e-Value) required for deciphering the homologies between pairs of genes and eventually constructing the homology matrix is set.

For each one of the n genomes that participate in mapID, containing $k_i$ ($i = 1 \ldots n$) genes each, the algorithm retrieves specific gene data such as the gene identifiers provided by KEGG, the respective Enzyme Commission (EC) identifiers and the gene sequences in FASTA format. This information is consequently used to construct n databases that can be used as reference gene data to apply the BLAST algorithm to. In the next step, $\sum_{i=1}^{n} (k_i \cdot k_f)$ blast searches are performed, where $k_f = \sum_{i=1}^{n} k_i$, and a matrix $C[k_f][k_f]$ is created containing the e-Values for all possible pairs of the participating genes. This leads eventually to a matrix $P[k_f][n]$ containing the phylogenetic profiles of all genes participating in the study. Phylogenetic profiles are binary representations that
record the presence or absence of a gene across a range of species \[19\]. The homology between genes is determined using the default BLAST settings and the e-Value threshold set as an input parameter (default value: $10^{-5}$). The data in this homology matrix $P$ are then clustered using the selected algorithm and a custom similarity metric based on the Jaccard metric. This custom metric extends the functionality of the Jaccard metric by appending to the numerator part the total number of attributes (in this case of genomes) where both genes whose similarity is evaluated do not have a homologue. Thus, the custom similarity metric is given by the following equation:

$$\text{sim}_{\text{jac}}(g_1, g_2) = \frac{m_{11} + m_{00}}{m_{01} + m_{10} + m_{11}} \quad (1)$$

where:

- $m_{11}$ represents the number of genomes where both genes ($g_1$ and $g_2$) have a homologue,
- $m_{00}$ represents the number of genomes where both genes ($g_1$ and $g_2$) do not have a homologue and
- $m_{10}$ ($m_{01}$) represents the number of genomes where gene $g_1$ ($g_2$) has a homologue while gene $g_2$ ($g_1$) does not.

Following that step, a validation of the clustering process is performed to determine the similarity and consequently the significance of the clusters generated utilizing the FASTA files for each of the clusters retrieved from KEGG. The homology score for each pair of clusters $i$ and $j$ (at index $i$, $j$ of the homology matrix) is the total number of homologues found between clusters $i$ and $j$ divided by the product of the total numbers of genes in clusters $i$ and $j$. The homology between each pair of genes is determined by executing blast searches for all pairs of FASTA sequences corresponding to the genomes participating in the current test case and considering that an e-Value score lower than $10^{-5}$ indicates the existence of homology between the examined genes. The similarity score for each pair of clusters $i$ and $j$ is the sum of the similarities (Eq. 1) between all pairs of genes in clusters $i$ and $j$ divided by the product of the total numbers of genes in clusters $i$ and $j$. The validation process is performed by comparing the intra-cluster similarity and homology to the respective inter-cluster values.

In the next step, the phylogenetic distances between the genomes that participate in the study are calculated using information derived from the clustering results and also from the topology of the metabolic pathway under examination. For this purpose, some custom lists of genes that satisfy three basic criteria are created. Specifically, each constructed distinct list contains genes that:

a) have been assigned to the same cluster
b) correspond to the same EC number
c) have the same KEGG Orthology Identifier (KO ID)

By taking all possible combinations of genes within each list, such that every genome is represented by only one gene at each combination, we calculate eventually the average phylogenetic distances of the genomes as they are anticipated by the sequence information carried out by the genes of the list.
Furthermore, each list extracted based on the aforementioned criteria contains genes either from all \( n \) genomes or, in case that is not possible, from a subset \( m < n, m > 2 \) of all genomes. In the first case, our method leads to the construction of a full phylogenetic tree while in the latter case we can create a sub-tree of the phylogenetic tree which can still yield some valuable results regarding the phylogenetic relationships between the organisms included. In every case, each tree constructed from a list conforming to the criteria (a), (b) and (c) is uniquely identified by a triplet of identifiers that characterizes this list, namely Cluster ID, EC Number and KO ID.

However, it is important to notice that lists with different triplets of identifiers, as described above, may lead to the same phylogenetic tree. For this reason, after the construction of all possible trees we examine if there are identical trees among the entire set of the retrieved trees. Then, each set of identical trees is assigned to a special category, or a Group. Thus, in the end we get a filtered version of the extracted trees so that each different tree structure is represented only once. Moreover, the conception of the Group categorization leads to a further classification of the proteins, besides the one achieved through the clustering that has been performed. Specifically, each EC Number that has been used in a list that conforms to the criteria (a), (b) and (c) is considered to belong to the Group that the tree of this list has been assigned to. Thus, the genes that correspond to this EC Number also receive a Group membership that constitutes a finer level of genes grouping extending the clustering information derived at an earlier step of the methodology. Finally, it should be noted that this specific method of phylogenetic trees construction leads eventually to the construction of zero, one or more phylogenetic trees (full tree, sub-trees or trees from both categories).

In the last step of the methodology, a post-processing procedure is applied to the data of the clusters in order to extract additional information regarding the composition of the clusters and the distribution of similar genes across the genomes. The acquired results are used to retrieve from KEGG the customized colored versions of the examined metabolic pathways by cluster and by genome. Those colored versions of the metabolic pathways contain special color annotation for the extracted clusters and Groups of genes. Specifically, each cluster that has been created corresponds to a distinct color that is used in the background space of each EC Number box in the pathway image. On the other hand, the assignments of genes into Groups is demonstrated by the color of the outline and of the numbers within the EC Number box. The colors selected for the clusters and Groups annotation do not indicate directly by themselves the specific clustering or allocation in Groups that is retrieved. Instead, distinction between clusters and Groups is achieved only by the contrast in colors used for the annotation. As a result, in order to assure that distinction between clusters and Groups is feasible only by visual inspection of the retrieved images, the colors are defined in our methodology in a way such that they provide sufficient diversity and contrast between them. The only exception is when an EC number contains genes that have been allocated to more than one clusters and this is denoted by a black color in the background space of the EC Number box. Besides, in the off-case an EC Number is not existent in any list used for the phylogenetic distances calculations, the numbers and outline color of the corresponding EC Number box in the pathway image are left white. Finally, the customized colored images of the pathways
retrieved from KEGG are stored locally and are available for visual evaluation and comparison.

The flowchart of the algorithm is shown in Figure 1.

3.1. Clustering algorithm evaluation

After rigorous experimentation using both the MCL and the EM clustering algorithms, we have concluded that the clusters generated with MCL are more robust and more closely correlated to the biological aspect of the problem, while the clusters produced by EM cannot be readily interpreted. Specifically, we present here the results obtained after applying the MCL and the EM clustering algorithms as part of the overall methodology to the Homo sapiens, the Arabidopsis thaliana and the Escherichia Coli K-12 MG1655 genomes with reference to the Glycolysis / Gluconeogenesis pathway. Both the two algorithms generate the same number of clusters and let \( \text{Cl}_{\text{MCL}} \) and \( \text{Cl}_{\text{EM}} \) be a cluster derived using MCL or EM respectively. In order to compare the \( \text{Cl}_{\text{MCL}} \) and \( \text{Cl}_{\text{EM}} \) clusters we choose first the cluster with the greatest number of nodes (\( n \)) in its graph and let \( \text{Cl}_{\text{MCL}} \) be that cluster. Then, the difference between \( \text{Cl}_{\text{MCL}} \) and \( \text{Cl}_{\text{EM}} \) is the number (\( m \)) of the edges that need to be added and/or deleted from \( \text{Cl}_{\text{EM}} \) in order to form \( \text{Cl}_{\text{MCL}} \) cluster. Thus, the similarity between those two clusters, expressed in a percent form, is given by the following equation:

\[
\text{similarity}_{\text{MCL}||\text{EM}} = (1 - \frac{m}{n}) \cdot 100\%
\]

After calculating the similarities between the clusters derived from MCL and EM (Table 1) we observed that there is in average over 85% similarity between the most similar clusters, as they are given by Eq. 2. However, those two clustering algorithms do not provide an equivalent level of granularity regarding the resulted clustering. Moreover, in all the experimental setups that were performed, EM cannot attain the distinction in separate groups of the genes belonging to a specific genome from the entire dataset, something that is achieved by MCL [20]. On the other hand, in sharp contrast with EM, the MCL algorithm leads to a degenerate clustering (i.e. singleton cluster) when the organisms selected for a test case have a close phylogenetic relationship. Thus, in case resulting clusters from EM can be sufficiently interpreted from a biological aspect, then we can assume that EM is more advantageous than MCL in cases of limited phylogenetic diversity. For the current study however, the development of the overall methodology is based on the use of MCL as the gene clustering algorithm.

Furthermore, the MCL algorithm’s efficiency and performance is significantly affected by the value assigned to the inflation parameter at the beginning of the clustering process [21]. Thus, we have determined through extensive experimentation the optimal value for the inflation parameter (‘12’) as the recommended value for the execution of the methodology (Table 2). Specifically, setting the inflation parameter to the optimal value stabilizes the number of clusters produced by the algorithm while at the same time yielding the optimal results (maximum values) with regards to the average intra-cluster similarity and homology.

Finally, the results of the clustering process are evaluated using the homology and similarity scores, which were defined earlier in section 3. The validity of the clusters is examined by comparing the homology and similarity scores calculated within the clusters to those calculated among the clusters. In all cases (some representative data shown in section 4), it was confirmed that the ratio of intra-cluster to inter-cluster values was in the
range of 10, with the notable exception of a few minor clusters were the ratio was closer to 1.

Table 1. Similarity (Eq. 2) between the clusters derived using either the MCL or the EM algorithm when the methodology is applied to the Homo sapiens, Arabidopsis thaliana and Escherichia Coli K-12 MG1655 genomes with reference to the Glycolysis / Gluconeogenesis pathway. Each percentage value is followed by a pair of values in [a+, b-] format where a and b correspond to the numbers of nodes that need to be added or deleted respectively from the cluster with the fewer nodes in order to form the cluster with the greatest number of nodes.

<table>
<thead>
<tr>
<th>Clusters derived from EM</th>
<th>Clusters derived from MCL</th>
<th>Similarity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>90 [2+, 0-]</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>60 [5+, 2-]</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>100 [0+, 0-]</td>
</tr>
<tr>
<td>Multi Cluster</td>
<td>Multi Cluster</td>
<td>100 [0+, 0-]</td>
</tr>
</tbody>
</table>

Table 2. Statistical overview of several different runs of the methodology using MCL for various values of the inflation parameter. The similarity metric used for this evaluation is given by Eq. (1). The reference pathway is the Glycolysis / Gluconeogenesis and the examined organisms are Escherichia Coli K-12 MG1655 (eco), Arabidopsis Thaliana (ath) και Homo Sapiens (hsa).

<table>
<thead>
<tr>
<th>MCL inflation</th>
<th>2</th>
<th>6</th>
<th>10</th>
<th>12</th>
<th>16</th>
<th>20</th>
<th>24</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>Results</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of clusters</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Average inter-cluster similarity</td>
<td>0.29</td>
<td>0.42</td>
<td>0.42</td>
<td>0.48</td>
<td>0.48</td>
<td>0.48</td>
<td>0.29</td>
<td>-</td>
</tr>
<tr>
<td>Standard deviation of inter-cluster similarity</td>
<td>0.21</td>
<td>0.21</td>
<td>0.28</td>
<td>0.28</td>
<td>0.28</td>
<td>0</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Average intra-cluster similarity</td>
<td>1.92</td>
<td>1.39</td>
<td>1.39</td>
<td>1.81</td>
<td>1.81</td>
<td>1.81</td>
<td>1.92</td>
<td>0.77</td>
</tr>
<tr>
<td>Standard deviation of intra-cluster similarity</td>
<td>1.53</td>
<td>0.41</td>
<td>0.41</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
<td>1.53</td>
<td>0</td>
</tr>
<tr>
<td>Average inter-cluster homology</td>
<td>0.035</td>
<td>0.02</td>
<td>0.02</td>
<td>0.016</td>
<td>0.016</td>
<td>0.016</td>
<td>0.035</td>
<td>-</td>
</tr>
<tr>
<td>Standard deviation of inter-cluster homology</td>
<td>0.035</td>
<td>0.017</td>
<td>0.017</td>
<td>0.014</td>
<td>0.014</td>
<td>0.014</td>
<td>0.035</td>
<td>-</td>
</tr>
<tr>
<td>Average intra-cluster homology</td>
<td>0.12</td>
<td>0.14</td>
<td>0.14</td>
<td>0.18</td>
<td>0.18</td>
<td>0.18</td>
<td>0.12</td>
<td>0.08</td>
</tr>
<tr>
<td>Standard deviation of intra-cluster homology</td>
<td>0.045</td>
<td>0.018</td>
<td>0.018</td>
<td>0.07</td>
<td>0.07</td>
<td>0.07</td>
<td>0.045</td>
<td>0</td>
</tr>
</tbody>
</table>
3.2. Algorithm complexity and efficiency

The maximum data size processed using that algorithm involved 325 genomes and 10,327 genes, due to memory space limits, considering 4 GBs of available memory space. The corresponding execution time of the algorithm reached about 15 hours using a 3.0 GHz Quad-Core Processor (Fig. 2).

![Fig. 2](image)

Fig. 2. Execution of the algorithm for different number of genes [a] and genomes [b]. In both cases the exponential character of the algorithm is evident.

3.3. Software implementation

The implementation of the algorithm outlined in the paper incorporates several different software tools and libraries that are combined and interconnected on top of a Java-based application framework. Specifically, data retrieval from KEGG database is achieved through SOAP-based web-services provided by the KEGG API. The construction of the blastable databases and the consequent BLAST searches are performed using the BLAST libraries (version 2.2.25+) provided by NCBI, whereas the output of the BLAST runs is read and organized using a custom XML DOM parser. The gene clustering is performed using the MCL implementation provided by the author of the original paper [22] while the EM’s implementation used is the one provided by WEKA Suite [23]. Finally, the application makes use of certain functions derived from the Matlab Bioinformatics toolbox [24] for the calculation of phylogenetic profiles using metabolic data in conjunction with information extracted from the resulted gene clustering. Those functions are accessed via Matlab generated Java components that are executed with MCR runtime engine [25]. The overall application is built on top of a common Java layer which is, besides the combination and coordination of the aforementioned heterogeneous tools and libraries, responsible for several pre/post-processing, evaluation and data mining tasks. Finally, the java tool that has been implemented is publicly available from the following URL: http://olympus.ee.auth.gr/~fpsom/alignPaths-pkg_Bundle_v0.9.tar.gz
4. Results overview

The proposed method was applied on several different pathways and genome sets and validated through both statistical methods and literature reviews. Two characteristic test cases are presented here, namely the application of the method on the Glycolysis / Gluconeogenesis metabolic pathway (KEGG identifier map00010) and on the Galactose Metabolism pathway (KEGG identifier map00052), which are both well known and extensively documented pathways [26].

The three genomes participating in the first test case are Arabidopsis Thaliana (for brevity denoted as ath), Escherichia Coli K-12 MG1655 (for brevity denoted as eco) and Homo Sapiens (for brevity denoted as hsa), and were selected for sufficient phylogenetic diversity. The total number of genes in this dataset is 209 (specifically ath: 105, eco: 39 and hsa: 65). The genomes used for the second test case are Arabidopsis Thaliana (ath), Drosophila melanogaster - fruit fly (for brevity denoted as dme) and Homo Sapiens (hsa) with a total number of genes at 103 (specifically ath: 44, dme: 30 and hsa: 29). The selected algorithm for both cases was MCL and the results are presented in detail in the following section.

4.1. Test cases

The results of the two test cases are presented here along with some interesting observations. Specifically, the pathway images cited below depict the distribution of the genes both by cluster and by genome after the clustering is performed and provide useful information about the internal structure of a single pathway as a result of genomes evolution.
4.1.1. **Case study – A**
Pathway: Glycolysis / Gluconeogenesis
Genomes: ath, eco, hsa

In each pathway image (Fig. 3) the EC identifiers of the corresponding genome are highlighted according to the cluster and the Group their genes belong to. A special case is the EC identifiers highlighted in black: they contain genes from more than one cluster. However, it must be noted that each gene is assigned to a single cluster, whereas an EC identifier, corresponding to several genes, may in turn belong to different clusters.

Figure 4 shows the distribution of the gene clusters on the pathway itself. It is interesting to note that Cluster 1 contains EC numbers corresponding to genes that constitute the main process of the pathway, whereas the EC identifiers of the fourth cluster contain genes only from the human genome. The second cluster contains genes only from Homo Sapiens and Arabidopsis Thaliana, as opposed to Cluster 3 that contains genes only from Escherichia Coli and Arabidopsis Thaliana. Finally, there exist several cases where an EC identifier corresponds to genes that individually belong to different clusters. These cases are shown collectively as a fifth cluster, but are also highlighted in each cluster diagram when the EC identifier contains at least one gene of the specific cluster.
Multi-genome core pathway identification through gene clustering

Cluster 1

Cluster 2

Cluster 3

Cluster 4
In the matter of validation, the produced clusters were evaluated using both the modified Jaccard similarity (Eq. 1) and the gene homology metrics. The intra-cluster similarity ($\approx 1.183$) was significantly higher than the corresponding inter-cluster similarity ($\approx 0.479$) (as shown in Figure 5). Moreover, the average number of homologies within the clusters ($\approx 0.182$) is also significantly higher than the corresponding value that is calculated between the clusters ($\approx 0.016$) thus indicating both the statistical significance and the validity of the clusters produced through our methodology.

In the current test case, 3 different kinds of trees have been extracted from 15 distinct lists of genes that follow the pattern (common Cluster ID, common EC Number, common KO ID), all belonging to cluster 1. The distribution of the 15 lists of genes that satisfy the phylogenetic trees construction criteria among those 3 unique Groups is shown in Figure 6 and the actual phylogenetic trees are shown in Figure 7. A close observation of these figures reveals that the correct phylogenetic tree corresponds to the tree of Figure 7.b as it separates the organisms into eukaryotes and non-eukaryotes (also confirmed by NCBI Taxonomy). It must be noted that there is not a globally proved process within our method for the automated extraction of the correct tree among the entire dataset of the constructed trees. However, in the current case study, the correct tree corresponds to the Group of genes containing the largest percentage of lists of genes.
Multi-genome core pathway identification through gene clustering

Fig. 5. Inter-cluster and intra-cluster homology [a] and similarity (Eq. 1) [b] figure with regard to test case A.

Fig. 6. Distribution of the lists of genes in the Groups that were formed for the calculation of the phylogenetic distances between the participating organisms with regard to test case A.
4.1.2. Case study - B

Pathway: Galactose metabolism  
Genomes: ath, dme, hsa

Although the Galactose metabolism pathway is relatively smaller than the Glycolysis/Gluconeogenesis pathway, we observe that the results are quite similar to the ones obtained in case study A. Initially, Figure 8 shows the distribution of the gene clusters among all participating genomes of the case study. One can easily observe in the separate pathway images by cluster (Fig. 9) that also in this case study, Cluster 1 contains EC numbers along the main chain of reactions of the pathway. Moreover, the EC identifiers of the third cluster contain genes only from the Arabidopsis Thaliana whereas the fourth cluster contain genes only from the human genome. Finally, the second cluster contains genes only from Homo Sapiens and Drosophila melanogaster.

In the matter of validation, the produced clusters were evaluated using both the modified Jaccard similarity (Eq. 1) and the gene homology metrics. The intra-cluster similarity and homology (≈ 2.289 and 0.426, respectively) were significantly higher than the corresponding inter-cluster values (≈ 0.4305 and 0.085, respectively as shown in Figure 10) denoting the significance and validity of the produced clusters.
Multi-genome core pathway identification through gene clustering

Homo Sapiens

Fig. 8. The Galactose metabolism pathway for the three genomes in the case study [B].

Cluster 1

Cluster 2

Cluster 3

Cluster 4

Multi cluster

Fig. 9. The Galactose metabolism pathway for each of the four produced clusters, and the case of EC identifiers with genes from multiple clusters (highlighted in black).
In this test case, there have been extracted 3 different kinds of trees from 9 distinct lists of genes that follow the pattern (common Cluster ID, common EC Number, common KO ID), all belonging to cluster 1. So, 3 distinct Groups of genes are formed in the end and the distribution of the 9 lists of genes among those Groups is shown in Figure 11. Besides, the images of the constructed phylogenetic trees are shown in Figure 12. We can notice through observing these images that the correct phylogenetic tree (confirmed by NCBI Taxonomy) corresponds to the tree of the Figure 11.a. Similarly with case study A, the correct phylogenetic tree in this case was extracted from the Group of genes that contains the largest percentage of lists of genes that have been used for the calculation of the phylogenetic distances among the participating genomes. Those preliminary results have also been identified in a large amount of various other test cases and constitute a strong indication that it is possible to construct the correct phylogenetic trees with a fully or partially automatic way through the novel method that we have developed.
Multi-genome core pathway identification through gene clustering

Fig. 12. Phylogenetic trees extracted from the Groups 1 (a), 2 (b) and 3 (c) with regard to test case B. (a) Tree structure: ((Homo Sapiens, Drosophila melanogaster), Arabidopsis Thaliana). (b) Tree structure: (Homo sapiens, (Drosophila melanogaster, Arabidopsis Thaliana)). (c) Tree structure: ((Homo sapiens, Arabidopsis Thaliana), Drosophila melanogaster).

5. Discussion

Attempting to identify the core pathway genes across a number of selected genomes, a novel methodology has been presented which utilizes the clusters produced by existing algorithms and infers their phylogenetic distance by means of a modified similarity metric. Although any clustering algorithm can be used in the process, the MCL and the EM algorithms have been used in the validation process, as they are better suited for sequence data. The output of the overall methodology can be easily utilized for the quick identification of essential sub-pathways of selected metabolic networks, and can infer in carefully selected sets of genomes, the evolutionary history of the different parts of a metabolic network.

As a driving force of pathway evolution is the introduction or removal of metabolites within the metabolic network [4], most approaches in literature aim to identify the phylogenetic relationships of the species within the properties of the metabolic pathway topologies [3]. This is usually achieved either through the pathway reaction content or through the analysis of the graph representation of the pathway. The methodology presented here is focused on the sequence data underlying the pathway organization, and attempts to project the inferred evolutionary relationships as an additional level of organization on the pathway network.
After a careful assessment of the results shown in the previous section, some interesting observations can be made. The presented test cases were among several different experimental setups and their results are representative for the performance of our methodology. First of all, in all cases and having considered sufficient phylogenetic diversity among the participating organisms, the first cluster always contained EC identifiers along the main reaction chain of the pathway, leading to the tentative conclusion that it may correspond to the highly conserved genes. Furthermore, in many cases, it is possible to extract groups of genes belonging only to specific genomes among the entire dataset. However, the algorithm is not efficient when used for species with close phylogenetic relationship (e.g. Homo sapiens, Pan troglodytes and Pongo abelii) in order to study highly conserved metabolic pathways in the respective species. This is a degenerate case as it leads to a single cluster for all the genes.

Moreover, the method suggested for the construction of the phylogenetic trees has been proven quite robust and reliable. In almost all cases, the correct tree corresponds to the Group of genes that contained the highest percentage of all lists of genes, specified in section 3. As a result, in most cases the correct tree can be extracted automatically without any need for manual curation or inspection of all the trees that are retrieved through our methodology.

Finally, by superimposing the highlighted pathway diagrams along the implied phylogenetic distance of the genomes, one may infer sub-chains of the pathway that have been transformed or evolved across the species. Besides, a thorough investigation of this problem, together with rigorous experimentation on several different and more elegantly selected sets of pathways/genomes may provide more information in these areas.

References